Generating non-verbal responses in virtual agent with use of LSTM network

Leon Koren and Tomislav Stipancic *

Posted Date: 15 January 2024

doi: 10.20944/preprints202401.1081.v1

Keywords: LSTM; facial expressions; framework; virtual agent; affective robotics

Article

Generating Non-Verbal Responses in Virtual Agent with Use of LSTM Network

Leon Koren [1] and Tomislav Stipancic [2,*]

[1] Faculty of mechanical engineering and naval architecture; leon.koren@fsb.unizg.hr
[2] Faculty of mechanical engineering and naval architecture; tomislav.stipancic@fsb.unizg.hr
* Correspondence: tomislav.stipancic@fsb.unizg.hr

Abstract: This paper investigates nonverbal communication in human interactions, with a specific focus on facial expressions. Employing a Long Short-Term Memory (LSTM) architecture and a customized facial expression framework, our approach aims to improve virtual agent interactions by incorporating subtle nonverbal cues. The paper contributes to the emerging field of facial expression generation, addressing gaps in current research and presenting a novel framework within Unreal Engine 5. The model's architecture, trained on the CANDOR corpus, captures temporal dynamics, and refines hyperparameters for optimal performance. During testing, the trained model showed a cosine similarity of -0.95. This enables the algorithm to accurately respond to non-verbal cues and interact with humans in a way that is comparable to human-human interaction. Unlike other approaches in the field of facial expression generation, the presented method is more comprehensive and enables the integration of a multi-modal approach for generating facial expressions. Future work involves integrating blendshape generation, real-world testing, and the inclusion of additional modalities to create a comprehensive framework for seamless human-agent interactions beyond facial expressions.

Keywords: LSTM; facial expressions; framework; virtual agent; affective robotics

1. Introduction

Human social interactions rely heavily on the interpretation of verbal and nonverbal cues to understand each other's states of mind [1]. While verbal cues are more straightforward to comprehend, nonverbal communication, encompassing facial expressions, body language, and even voice modulation devoid of linguistic features, plays a pivotal role in conveying a wealth of information about an individual's emotional and psychological state during interactions [1]. Micro expressions, involuntary facial expressions that betray underlying emotions, are particularly significant in this context. Emotions are an important ingredient of every social interaction [2]. A person's emotions are a strong decision-making factor that shapes their reasoning process. However, emotions are also highly subjective and can change as new information emerges. The process of perceiving emotions in emotion-aware systems is based on various sensing modalities, including visual, auditory, and tactile, as reported in [3]. Interestingly, research suggests that individuals tend to overestimate their proficiency [4].

This paper delves into the realm of nonverbal interaction, specifically focusing on facial expressions, to develop a virtual agent capable of meaningful interactions with humans. The chosen approach employs a Long Short-Term Memory (LSTM) architecture in tandem with a custom-built facial expression framework. This strategy leverages the established effectiveness of LSTM in generating nonverbal responses, which are then integrated into the virtual agent to enhance interaction quality and expedite the grounding process.

The implemented solution is rigorously tested on PLEA, an affective robot with a biomimicking interactive robot head capable of displaying facial expressions [5], as illustrated in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

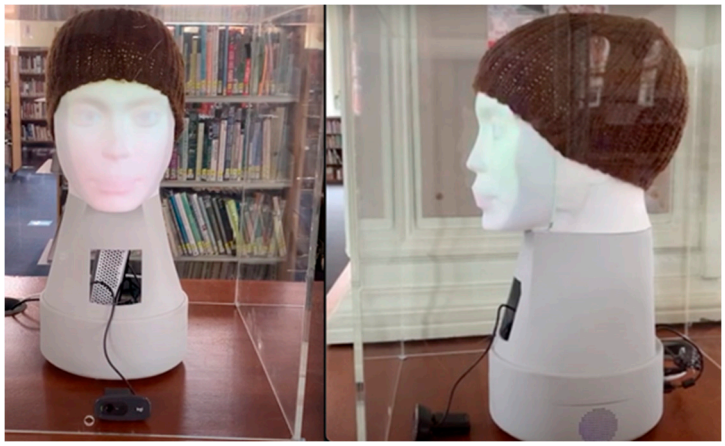

Figure 1. Interactive biomimicking robot head PLEA during the ART & AI Festival 2021, Leicester, UK.

The subsequent sections of this paper are organized as follows: Section 2 provides an overview of the current state of the art in autonomous expression generation and facial expression recognition, identifying similar concepts. It also explains the contributions of the proposed methodology to the existing body of knowledge. Section 3 introduces the framework for autonomous expression generation, while Section 4 details the datasets used, necessary adaptations, and data filtering procedures, emphasizing simplification and normalization techniques. Section 5 delves into the algorithm for expression generation, the implementation of the LSTM architecture, training results, and techniques employed for hyperparameter tuning. The evaluation of the algorithm, involving interactions with PLEA and feedback from participants through standardized questionnaires, is discussed in Section 6. Finally, Section 7 concludes the paper by summarizing key concepts and suggesting potential avenues for future research.

2. State of the art

Facial expressions are nonverbal signals that can be used to indicate one's status in a conversation, e.g., via backchannelling or rapport, as reported in [6,7]. In [8], authors presented a chatbot conversational agent that relies on textual analysis to assist older adults' well-being. This functionality is integrated into various similar software applications. For example, in [9] authors provided an overview of the use of Social Robots in Mental Health and Well-Being Research. In the field of facial expression research, while facial expression recognition has become mainstream, the development of facial expression generation tasks is just starting to gain traction. This is fueled by advancements in human-like 3D models such as Unreal Engine's Metahuman and Unity Reallusion. Currently, there's a lack of established state-of-the-art algorithms for autonomous facial expression generation, presenting a unique opportunity for researchers to contribute to this emerging area.

A significant contribution to the field of facial expression generation is the work of Yu et al. [10]. In their research, the authors propose a new approach that considers two crucial factors for realistic facial animation: believable visual effects and valid facial movements. They mainly focus on defining valid facial expressions and generating facial movements in a human-like manner. They achieved this through the use of facial expression mapping based on local geometry and with the addition of dynamic parameters based on psychophysical methods.

Another contribution is the FExGAN-Meta dataset and model [11], specifically designed for generating facial expressions in metahumans. The authors not only introduced a Facial Expression Generation model but also created a substantial dataset of metahuman facial images with corresponding expression labels. This dataset enables the trained model to generate various facial expressions on any given face using only an input image and an array specifying the desired expression. In contrast, our work focuses on autonomously generating facial expression vectors based on prior knowledge.

 doi:10.20944/preprints202401.1081.v1

In contrast to other state-of-the-art approaches, Otberdout et. al. [12] presented facial expressions generation system based on Hilbert Hypersphere and Generative Adversarial Networks. In their work authors synthesize facial expressions from neutral facial expression images. Authors can generate images and/or videos of six different facial expressions. In contrast, this paper uses abstract concepts of facial expression to generate new expressions based on previous knowledge.

Within the realm of facial expression recognition, recent advancements have been driven by deep learning approaches. Kue et al. [13] introduced a method for facial expression recognition, rigorously evaluated across multiple image datasets. Khaireddin et al. [14] proposed a CNN-based architecture that achieved high accuracy on the FER2013 dataset through careful parameter optimization. Additionally, Sajjanhar et al. [15] presented and evaluated another CNN-based architecture for facial expression recognition, leveraging a pre-trained face model [16].

In the domain of generative adversarial networks (GANs), Deng et al. [17] developed a GAN-based network capable of projecting images into a latent space. They then used a Generalized Linear Model (GLM) to capture the directional aspects of various facial expressions. Expanding on GAN architectures, Deng et al. [18] proposed a GAN architecture capable of both classifying and generating facial expression images. Furthermore, [19] presented a disentangled GAN network-based architecture proficient in generating images with user-specified facial expressions.

These contributions collectively highlight the growing interest and diverse approaches within the field, shedding light on both the potential and the current gaps in the realm of facial expression generation and recognition.

3. Framework

The facial expression generation framework comprises of dedicated modules, as depicted in Figure 3. Module *Input picture* (1) grabs an image from the video stream and finds all faces on it. Module chooses a face based on the algorithm described in Koren et.al [20], crops it and resizes it to a predetermined size. *CNN facial expression extraction* (2) module takes previously created images and with the use of the efficient residual neural network (ENet) [21] extracts seven standard expressions in the form of an array. The algorithm was used as a standard network trained on a large dataset AffectNet [22].

The module numbered 3 (LSTM facial expression generation) is a novel contribution of this work. The arrows connecting the modules signify the transfer of data, with associated vector dimensions specified.

In addition, Unreal Engine 5 (UE5) serves as a pivotal component within the framework (5). Within the UE5 instance, facial expression visualization is achieved through the utilization of Metahuman technology [23]. This integration allows for realistic and dynamic facial expressions, enhancing the overall immersive experience.

Figure 2. shows the process of facial expression generation based on the presented framework. The diagram shows how the virtual agent expresses different facial expression from real person. In this way, virtual agent is not just mimicking facial expression of real person but generating response based on learned interactions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

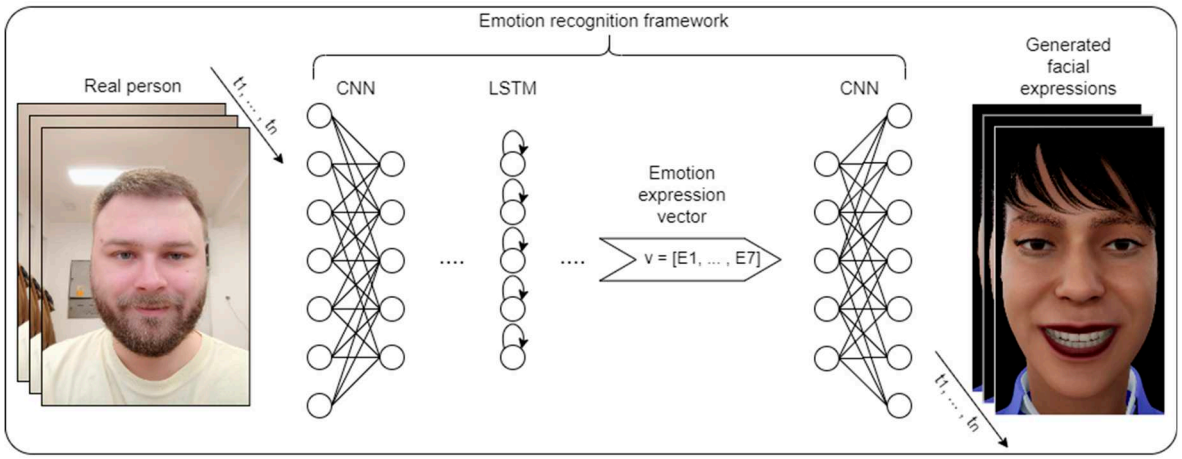

Figure 2. Process of facial expression generation.

Modules one, two and five represent those that have already been implemented, whereas module four is scheduled for future development.

Moreover, UE5 facilitates interaction with remote users via the Web Real-Time Communication (WebRTC) protocol. This feature enables seamless communication and collaboration, as users can remotely engage with the virtual entity generated by the framework. The incorporation of WebRTC protocol ensures low latency and high-quality communication, contributing to a more natural and engaging user experience.

Furthermore, the entire framework operates as a single instance on a central workstation. This design choice not only optimizes computational resources but also provides the compelling effect of a cohesive virtual entity. The consolidation of the entire framework within the UE5 instance underscores its efficiency and cohesiveness, streamlining the generation and visualization of facial expressions.

The key features collectively position the facial expression generation framework as an integrated and robust system, with Unreal Engine 5 serving as the foundational platform for its immersive and interactive capabilities.

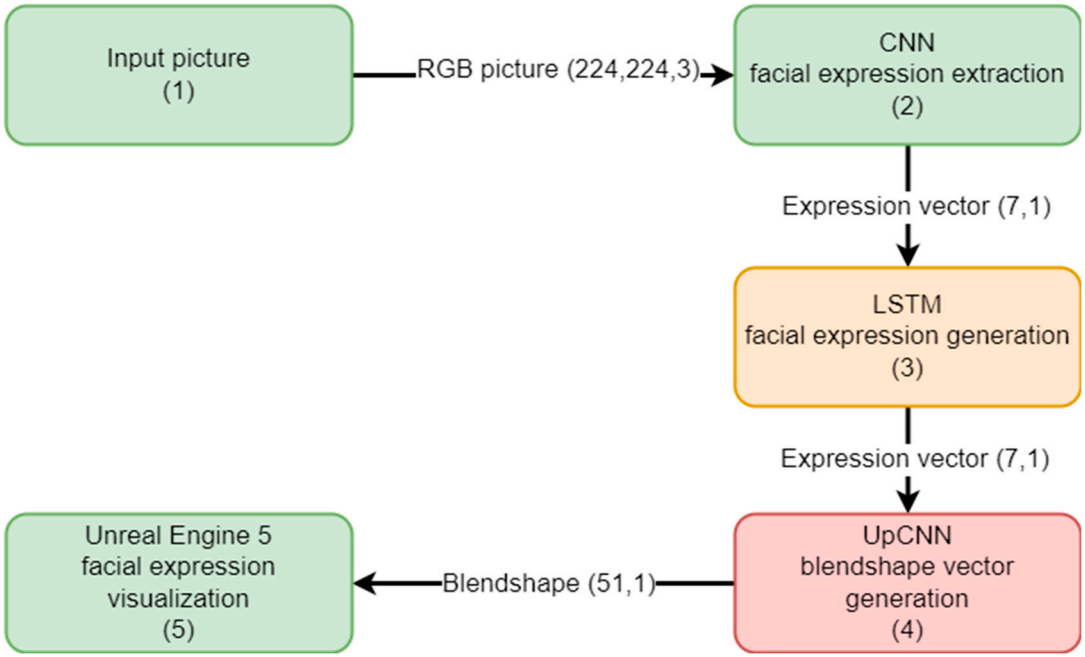

Figure 3. Facial expression generation framework.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

5. Facial expression generation model

The model architecture employed in this paper consists of a multi-layer Long Short-Term Memory (LSTM) network, as seen in Figure 4. The input to the model is represented as a vector with seven values, denoted as E:

$$E = [Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise]$$

The model utilizes a vector representation where each element is constrained within the range of zero to one, with the sum of all elements equating to one. This vector efficiently encapsulates emotional states, providing a quantitative framework for processing nuanced emotional information.

For temporal analysis, a 48-array time series is employed during training, introducing a dynamic dimension to the model. This allows for the capture of sequential dependencies and the identification of temporal patterns within the emotional data, enhancing the model's ability to discern evolving emotional states over consecutive data points.

To facilitate a seamless transition from the LSTM layers to the final output, a fully connected layer is incorporated. This layer utilizes a softmax activation function, ensuring that the output adheres to the same format as the input vector. The strategic integration of this fully connected layer is grounded in the principles of mathematical coherence and alignment with the inherent characteristics of the emotional state representation.

In the process of refining the model for optimal performance, an exhaustive exploration of hyperparameter configurations is undertaken. Hyperparameters include every parameter that needs to be defined before training has commenced [17]. The best example of this parameter is the learning rate, without the learning rate specified training is impossible. With different combinations of these parameters, values results can differ greatly. Figure 4 shows the architecture of the LSTM network before hyperparameter optimization.

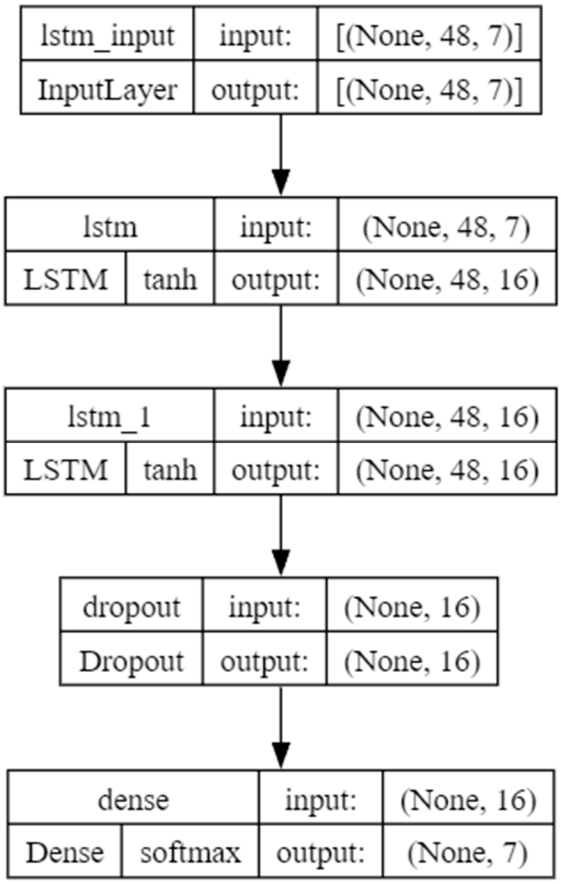

Figure 4. Architecture without hyperparameter optimization.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

For hyperparameter optimization, the authors used the following parameters: Learning rate, Number of units in LSTM layer, Number of LSTM layers and Dropout rate. Hyperparameter optimization can be done with many different algorithms [24]. For this optimization, the HyperBand [25] algorithm was used, as is the current state-of-the-art algorithm for hyperparameter optimization. After optimization, LSTM network parameters were as follows:

Table 1. Comparison of hyperparameters before and after optimization.

Hyperparameters	Before optimization	After optimization
Learning rate	1e-4	5e-4
Number of units in LSTM layer	16	40
Number of LSTM layers	1	2
Dropout rate	0.3	0.25

The Architecture of the LSTM network after optimization is shown in Figure 5.

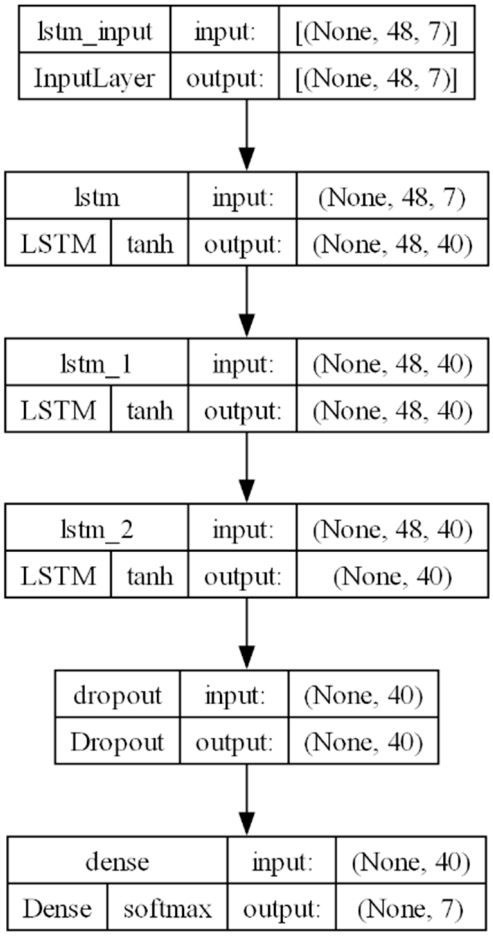

Figure 5. LSTM network architecture after hyperparameter optimization.

4. Dataset

The dataset utilized in this study is derived from the CANDOR corpus [26], encompassing 1656 English conversations conducted through video chat, with a cumulative video duration of approximately 850 hours. Notably, the video conversations were meticulously synchronized between the speakers' cameras.

The preprocessing of the dataset involved extracting video files and converting them into frames while maintaining synchronization. The OpenFACE library [27] facilitated the extraction and alignment of faces from each frame, ensuring synchronization throughout the process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

The subsequent phase focused on leveraging the extracted and aligned faces to discern facial expressions. The HSEmotion model [28], utilizing the ENet architecture, was applied for the extraction of facial expressions, capturing nuanced emotional expressions from the facial features in the frames.

To structure the dataset for analysis, the input data underwent segmentation into batches, each containing 48 timeframes, with each batch shifted forward by one timeframe. This meticulous segmentation and shifting strategy ensured comprehensive coverage of temporal dynamics within the dataset. The output data, representing the target variable, was configured to correspond to the timeframe situated one step into the future from the last input data timeframe in each batch.

The methodology encompassed the extraction and alignment of faces from synchronized video frames using the OpenFACE library. The HSEmotion model, specifically utilizing the ENet architecture, was then employed to extract facial expressions. The resultant dataset was systematically formatted, with input data organized into batches with forward shifts, facilitating subsequent predictive modelling, as illustrated in Figure 6. This systematic approach ensures a robust foundation for analyzing temporal patterns in facial expressions within the context of video conversations, contributing valuable insights to the study of interpersonal communication.

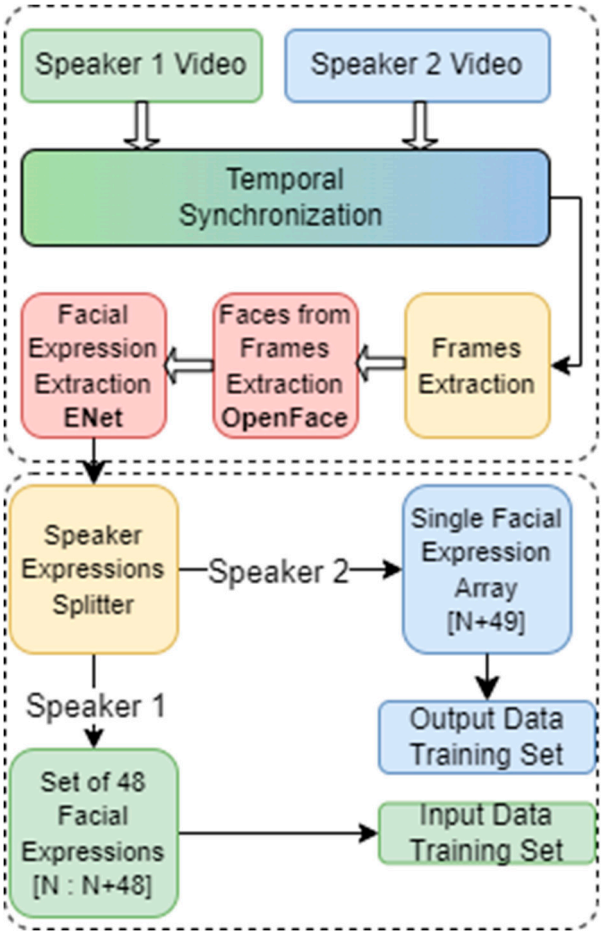

Figure 6. Flow chart of data preprocessing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

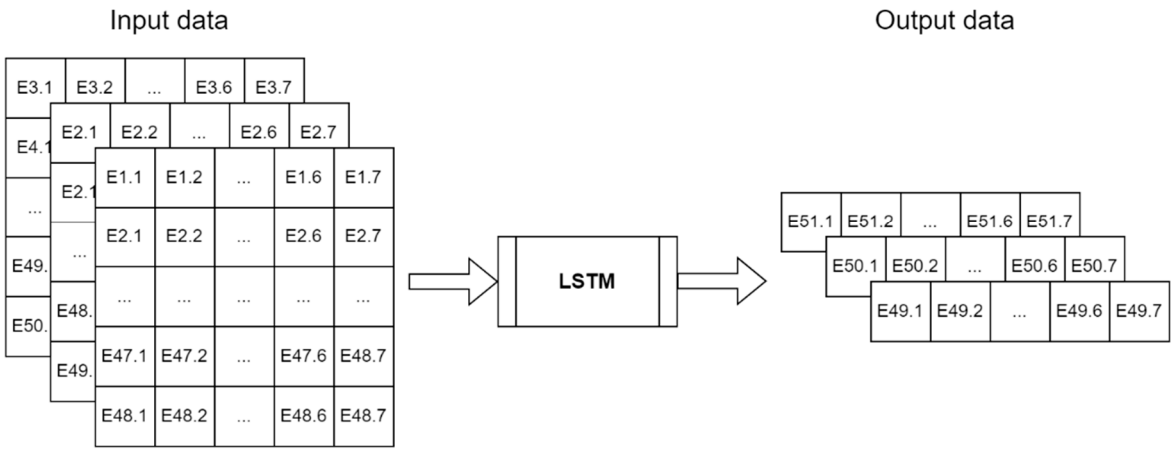

Figure 7. Graphical representation of training dataset.

6. Results

The training process consists of 60 epochs, employing a cosine similarity loss function and the Adam optimizer. Early stopping and model checkpoint callbacks are incorporated for enhanced efficiency. Early stopping prevents overfitting by terminating training when the model's validation set performance plateaus, while model checkpointing ensures the preservation of the best-performing model.

Figure 8. shows the validation and training loss of the model without hyperparameter optimization while Figure 9. shows the same data for a model with hyperparameter optimization.

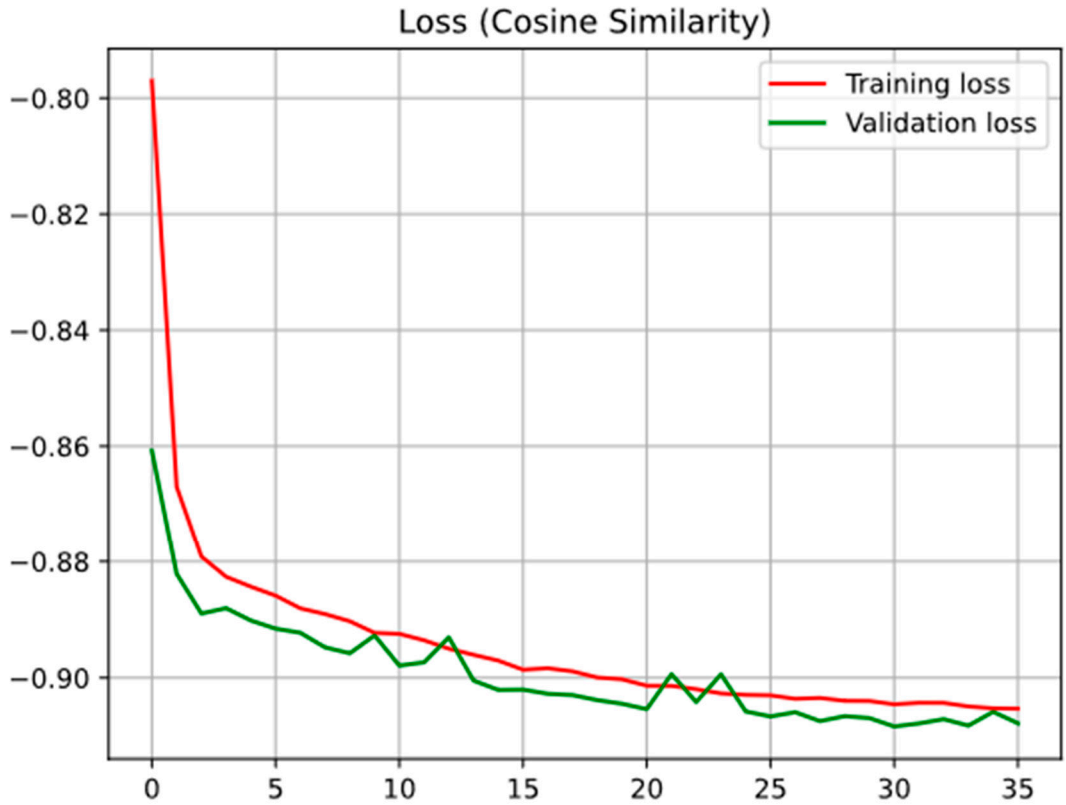

Figure 8. Training and validation loss without hyperparameter optimization.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1081.v1

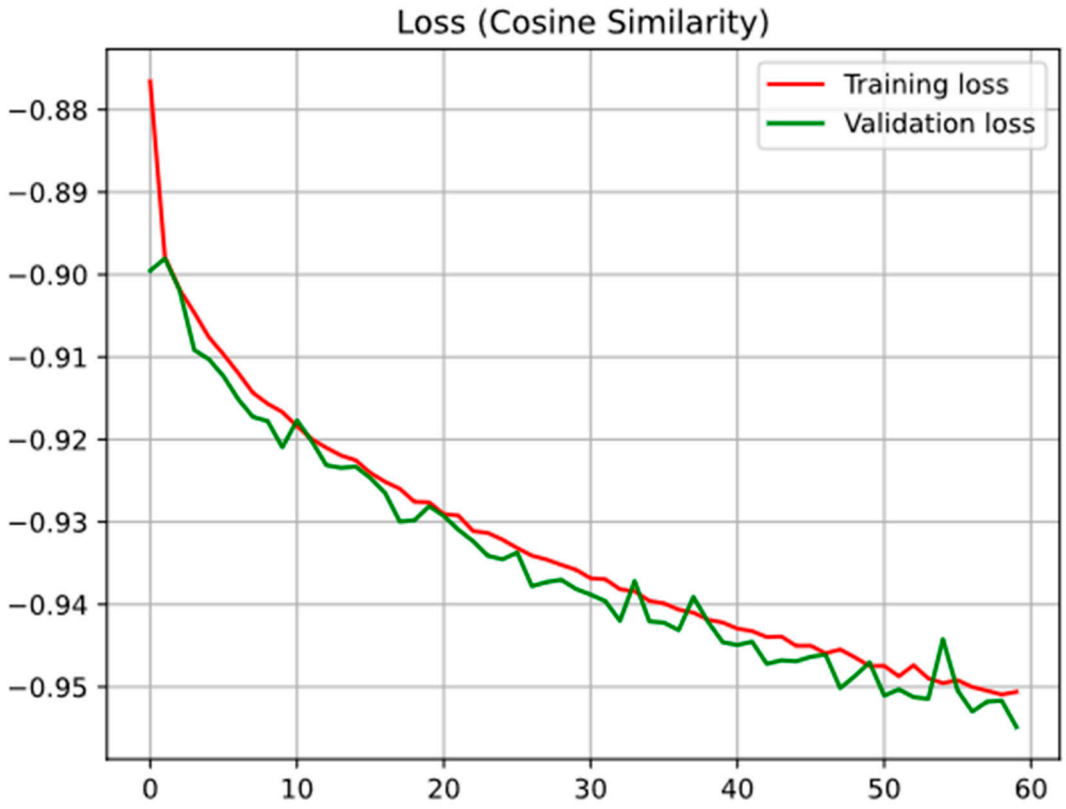

Figure 9. Training and validation loss with hyperparameter optimization.

The hyperparameter optimization model had better performance in terms of final loss, but due to the smaller learning rate model needed more epochs to achieve this performance. While Early stopping callback was implemented in both runs from graphs can be seen that for the second one early stopping was not employed (training stopped at 60 epochs) while at first one loss stopped improving at the 32nd epoch and callback stopped training at 35th epoch. In this way, overtraining was prevented with the additional benefit of saving computational power.

The cosine similarity loss utilized quantifies the similarity between two vectors, with 1 indicating greater dissimilarity and -1 indicating greater similarity. Notably, the model's loss approaches -1, suggesting that its predicted values closely align with the ground truth.

This systematic approach to hyperparameter tuning and the implementation of strategies like early stopping and model checkpointing contribute to the model's efficacy, reflected in the nearly -1 cosine similarity loss. The chosen hyperparameters and training strategies collectively establish a model that accurately and reliably aligns its predictions with the ground truth for the given task.

7. Future work

The presented model constitutes a fundamental component within a comprehensive facial expression generation framework. The subsequent phase involves the development of a dedicated model for the generation of blendshapes, serving as controls for the expressions exhibited by the virtual agent. Furthermore, the entire framework is slated for integration into the PLEA virtual agent, as briefly introduced earlier.

Following implementation, rigorous testing of the framework is planned, incorporating real human subjects across a spectrum of scenarios. Each scenario will be meticulously crafted to systematically assess specific facets of the framework's capacity to convey non-verbal communication to humans effectively. The evaluation of the interactions' quality will be conducted through the administration of standardized questionnaires.

As an integral part of the evaluation process, the framework's efficacy will be gauged by its performance in diverse real-world scenarios, ensuring its adaptability and reliability. This empirical

 doi:10.20944/preprints202401.1081.v1

approach aims to validate the framework's practical utility in facilitating seamless human-agent interactions.

To enhance the overall interaction quality and expedite the grounding process between the interacting human and virtual agent, additional modalities will be incorporated into the framework. The integration of these modalities is informed by the work of Koren et al. [29], providing a theoretical foundation for the augmentation of interaction modalities. The objective is to leverage supplementary channels of communication beyond facial expressions, thus fortifying the agent's ability to convey nuanced information and respond dynamically to human cues.

The ensuing phases of this research endeavor will center on a meticulous analysis of the introduced modalities and their impact on the overall interaction quality. This iterative process aligns with a commitment to continual improvement and refinement, aiming to establish a robust framework for human-agent interactions that extends beyond facial expressions to encompass a multi-modal communication paradigm.

8. Conclusion

In conclusion, this paper introduces a practical and technically grounded facial expression generation framework with significant implications for human-agent interaction. The utilization of an LSTM architecture, coupled with Unreal Engine 5 and the Web Real-Time Communication protocol, facilitates a tangible improvement in the virtual agent's capacity to engage with users through nuanced nonverbal cues. The model, meticulously trained on the CANDOR corpus, emerges as a robust tool with high accuracy and temporal sensitivity, showcasing its ability to capture the nuanced evolution of emotional states during interactions.

The systematic approach employed in hyperparameter tuning and model training, including the implementation of early stopping and model checkpointing, ensures the reliability and stability of the model. The achieved nearly perfect cosine similarity loss underscores the model's proficiency in aligning its predictions closely with the ground truth, further validating its effectiveness in generating facial expressions that mirror genuine human emotional responses.

Looking forward, the outlined roadmap for future work is geared towards enhancing the framework's versatility and real-world applicability. The integration of blendshape generation, planned real-world testing involving human subjects across diverse scenarios, and the incorporation of additional communication modalities mark the next steps. This evolution aims to establish a comprehensive and adaptive framework for human-agent interactions, transcending the realm of facial expressions to embrace a multi-modal communication paradigm. The use of a multimodal approach can enhance situational embodiment, self-explanatory nature, and context-driven interaction to increase interactivity [30].

The iterative and empirical nature of this research underscores a commitment to continuous improvement, with a focus on refining the framework to meet the demands of varied scenarios and interactions. By addressing existing gaps in facial expression research and pushing the boundaries of virtual interaction, this work contributes to the ongoing evolution of human-computer interaction. The potential impact extends beyond the realm of virtual agents, fostering advancements in fields where nuanced nonverbal communication plays a pivotal role. This research lays the foundation for a more nuanced, effective, and adaptable approach to human-agent collaboration, with practical applications in diverse settings.

Author Contributions: Conceptualization, L.K..; methodology, T.S. and L.K.; software, L.K.; validation, L.K. and T.S.; formal analysis, L.K.; investigation, L.K.; resources, L.K.; data curation, L.K.; writing—original draft preparation, T.S. and L.K; writing—review and editing, T.S.; visualization, L.K.; supervision, T.S.; project administration, T.S.; funding acquisition, T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was in part funded by Croatian Science Foundation, grant number UIP-2020-02-7184.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from BetterUp Labs and are available https://betterup-data-requests.herokuapp.com/ with the permission of dr. Gus Cooney.

Acknowledgments: This work has been supported in part by the Croatian Science Foundation under the project "Affective Multimodal Interaction based on Constructed Robot Cognition—AMICORC (UIP-2020-02-7184).".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hall, J.A.; Horgan, T.G.; Murphy, N.A. Nonverbal Communication. *Annu Rev Psychol* **2019**, *70*, 271–294. https://doi.org/10.1146/ANNUREV-PSYCH-010418-103145.
2. Perusquía-Hernández, M.; Jáuregui, D.A.G.; Cuberos-Balda, M.; Paez-Granados, D. Robot Mirroring: A Framework for Self-Tracking Feedback through Empathy with an Artificial Agent Representing the Self. **2019**.
3. Spitale, M.; Okamoto, S.; Gupta, M.; Xi, H.; Matarić, M.J. Socially Assistive Robots as Storytellers That Elicit Empathy. *ACM Transactions on Human-Robot Interaction (THRI)* **2022**, *11*. https://doi.org/10.1145/3538409.
4. Vrij, A.; Hartwig, M.; Granhag, P.A. Annual Review of Psychology Reading Lies: Nonverbal Communication and Deception. *Annu. Rev. Psychol* **2019**, *70*, 295–317. https://doi.org/10.1146/annurev-psych-010418.
5. Stipancic, T.; Koren, L.; Korade, D.; Rosenberg, D. PLEA: A Social Robot with Teaching and Interacting Capabilities. *Journal of Pacific Rim Psychology* **2021**, *15*. https://doi.org/10.1177/18344909211037019/ASSET/IMAGES/10.1177_18344909211037019-IMG1.PNG.
6. Bavelas, J.B.; Gerwing, J. The Listener as Addressee in Face-to-Face Dialogue. *International Journal of Listening* **2011**, *25*, 178–198. https://doi.org/10.1080/10904018.2010.508675.
7. Rawal, N.; Stock-Homburg, R.M. Facial Emotion Expressions in Human–Robot Interaction: A Survey. *Int J Soc Robot* **2022**, *14*, 1583–1604. https://doi.org/10.1007/S12369-022-00867-0/TABLES/5.
8. El Kamali, M.; Angelini, L.; Caon, M.; Khaled, O.A.; Mugellini, E.; Dulack, N.; Chamberlin, P.; Craig, C.; Andreoni, G. NESTORE: Mobile Chatbot and Tangible Vocal Assistant to Support Older Adults' Wellbeing. *ACM International Conference Proceeding Series* **2020**. https://doi.org/10.1145/3405755.3406167.
9. Scoglio, A.A.J.; Reilly, E.D.; Gorman, J.A.; Drebing, C.E. Use of Social Robots in Mental Health and Well-Being Research: Systematic Review. *J Med Internet Res* **2019**, *21*. https://doi.org/10.2196/13322.
10. Yu, H.; Garrod, O.; Jack, R.; Schyns, P. A Framework for Automatic and Perceptually Valid Facial Expression Generation. *Multimed Tools Appl* **2015**, *74*, 9427–9447. https://doi.org/10.1007/S11042-014-2125-9/FIGURES/12.
11. Siddiqui, J.R. FExGAN-Meta: Facial Expression Generation with Meta Humans. **2022**.
12. Otberdout, N.; Daoudi, M.; Kacem, A.; Ballihi, L.; Berretti, S. Dynamic Facial Expression Generation on Hilbert Hypersphere With Conditional Wasserstein Generative Adversarial Nets. *IEEE Trans Pattern Anal Mach Intell* **2022**, *44*, 848–863. https://doi.org/10.1109/TPAMI.2020.3002500.
13. Kuo, C.-M.; Lai, S.-H.; Sarkis, M. A Compact Deep Learning Model for Robust Facial Expression Recognition.
14. Khaireddin, Y.; Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. **2021**.
15. Sajjanhar, A.; Wu, Z.; Wen, Q. Deep Learning Models for Facial Expression Recognition. *International Conference on Digital Image Computing: Techniques and Applications* **2018**. https://doi.org/10.1109/DICTA.2018.8615843.
16. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. VGGFace2: A Dataset for Recognising Faces across Pose and Age. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018* **2018**, 67–74. https://doi.org/10.1109/FG.2018.00020.
17. Ning, X.; Xu, S.; Zong, Y.; Tian, W.; Sun, L.; Dong, X. Emotiongan: Facial Expression Synthesis Based on Pre-Trained Generator. *J Phys Conf Ser* **2020**, *1518*, 012031. https://doi.org/10.1088/1742-6596/1518/1/012031.
18. Deng, J.; Pang, G.; Zhang, Z.; Pang, Z.; Yang, H.; Yang, G. CGAN Based Facial Expression Recognition for Human-Robot Interaction. *IEEE Access* **2019**, *7*, 9848–9859. https://doi.org/10.1109/ACCESS.2019.2891668.
19. Ali, K.; Hughes, C.E. Facial Expression Recognition Using Disentangled Adversarial Learning. **2019**.
20. Koren, L.; Stipancic, T.; Ricko, A.; Orsag, L. Person Localization Model Based on a Fusion of Acoustic and Visual Inputs. *Electronics 2022, Vol. 11, Page 440* **2022**, *11*, 440. https://doi.org/10.3390/ELECTRONICS11030440.
21. Savchenko, A. V.; Savchenko, L. V.; Makarov, I. Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network. *IEEE Trans Affect Comput* **2022**, *13*, 2132–2143. https://doi.org/10.1109/TAFFC.2022.3188390.
22. Mollahosseini, A.; Member, S.; Hasani, B.; Mahoor, M.H.; Member, S. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild.
23. Fang, Z.; Cai, L.; Wang, G. MetaHuman Creator the Starting Point of the Metaverse. *Proceedings - 2021 International Symposium on Computer Technology and Information Science, ISCTIS 2021* **2021**, 154–157. https://doi.org/10.1109/ISCTIS51085.2021.00040.

24. Yang, L.; Shami, A. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing* **2020**, *415*, 295–316. https://doi.org/10.1016/J.NEUCOM.2020.07.061.
25. Li, L.; Jamieson, K.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research* **2018**, *18*, 1–52. https://doi.org/10.5555/3122009.
26. Reece, A.; Cooney, G.; Bull, P.; Chung, C.; Dawson, B.; Fitzpatrick, C.; Glazer, T.; Knox, D.; Liebscher, A.; Marin, S. The CANDOR Corpus: Insights from a Large Multimodal Dataset of Naturalistic Conversation. *Sci Adv* **2023**, *9*. https://doi.org/10.1126/SCIADV.ADF3197/SUPPL_FILE/SCIADV.ADF3197_SM.PDF.
27. Amos, B.; Ludwiczuk, B.; Satyanarayanan, M. OpenFace: A General-Purpose Face Recognition Library with Mobile Applications. **2016**.
28. Savchenko, A. V.; Savchenko, L. V.; Makarov, I. Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network. *IEEE Trans Affect Comput* **2022**, *13*, 2132–2143. https://doi.org/10.1109/TAFFC.2022.3188390.
29. Koren, L.; Stipancic, T.; Ricko, A.; Orsag, L. Multimodal Emotion Analysis Based on Visual, Acoustic and Linguistic Features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2022**, *13315 LNCS*, 318–331. https://doi.org/10.1007/978-3-031-05061-9_23.
30. Stipancic, T.; Jerbic, B.; Curkovic, P. Bayesian Approach to Robot Group Control. *Lecture Notes in Electrical Engineering* **2013**, *130 LNEE*, 109–119. https://doi.org/10.1007/978-1-4614-2317-1_9/COVER.